User Profiling through Deep Multimodal Fusion

Golnoosh Farnadi
University of California, Santa Cruz and Ghent University
gfarnadi@ucsc.edu

Jie Tang
Tsinghua University
jietang@tsinghua.edu.cn

Martine De Cock
University of Washington, Tacoma and Ghent University
mdecock@uw.edu

Marie-Francine Moens
KU Leuven
sien.moens@cs.kuleuven.be

ABSTRACT

User profiling in social media has gained a lot of attention due to its varied set of applications in advertising, marketing, recruiting, and law enforcement. Among the various techniques for user modeling, there is fairly limited work on how to merge multiple sources or modalities of user data – such as text, images, and relations – to arrive at more accurate user profiles. In this paper, we propose a deep learning approach that extracts and fuses information across different modalities. Our hybrid user profiling framework utilizes a shared representation between modalities to integrate three sources of data at the feature level, and combines the decision of separate networks that operate on each combination of data sources at the decision level. Our experimental results on more than 5K Facebook users demonstrate that our approach outperforms competing approaches for inferring age, gender and personality traits of social media users. We get highly accurate results with AUC values of more than 0.9 for the task of age prediction and 0.95 for the task of gender prediction.

KEYWORDS

User modeling, Age and gender prediction, Personality prediction, Social media, Deep neural networks

1 INTRODUCTION

Nowadays users actively generate content in many online social media platforms. User profiling by inferring users' age, gender and personality traits plays an important role in providing personalized services, viral marketing, recommender systems and tailored advertisements [17]. Previous work in the field of psychology has highlighted the value of identifying the personality traits of users as an aid in building adaptive and personalized systems to provide rich and improved user experiences [18, 27].

Various computational approaches of user profiling based on user-generated content (UGC) have been proposed in recent years [6, 22, 23]; more details on related works are presented in Section 2. Much of these efforts are aimed at finding novel techniques to infer user profiles using only one type of information, such as the user's textual posts. However, in many social media platforms, users generate content in different modalities, such as textual content (e.g., status updates, blog posts, tweets, comments, etc.) and visual content (e.g., photo and video), while also connecting with each other, i.e., creating relational content. A framework that leverages all available information about users can learn more accurate user profiles. This is especially useful for platforms where not every user generates the same type of information, and models trained based on one source of information fail to produce accurate user profiles. Examples include users who write status updates but never upload pictures, or users who join social media platforms only to consume knowledge and to relate with each other, rather than producing any textual or visual content themselves.

Neural networks lend themselves well for integrating multiple data sources, as they allow a non-linear combination of data sources to be trained to solve the problem. An early example is the use of a time-delay neural network (TDNN) to handle temporal multimodal data [28]. In general, neural networks are considered as a suitable technique to learn non-linear mappings in high-dimensional settings, however due to their slow training, they did not immediately take off as a popular technique for modeling multiple data sources. Recently, deep neural networks (DNNs), with the help of Graphics Processing Units (GPUs) which reduce the training time, gained a lot of attention. DNNs are arguably the best known method for most pattern recognition problems involving perception. They already perform on a human level on many important tasks such as handwritten digit recognition, pedestrian tracking, etc. One of the main advantages of DNN methods is that they do not need to rely on human designed features. Instead, DNNs learn their features from raw inputs. As a general reference to deep neural networks, we refer to [9].

In this paper, we make several contributions. First, we present a novel hybrid DNN based framework which we call "User Profiling through Deep Multimodal Fusion (UDMF)" that integrates multiple sources of user data for user profiling. We introduce a mechanism of stacking to leverage the dependency among target variables to more accurately infer user attributes. Second, we design a hybrid level of modeling multiple data sources with power-set combination. Using power-set combination, UDMF incorporates shared and non-shared representations among the data sources and integrates them both at the feature level and at the decision level. We perform user profiling using three modalities, namely textual data (i.e., users' posts), visual data (i.e., profile pictures) and relational data (i.e., users' page likes). For the social relational content in social media, our third contribution is that we propose to leverage a relational embedding approach called Node2Vec, in which we extract relational features from the social graph by performing a random walk through the graph. To the best of our knowledge this paper is the first paper that uses a Node2Vec embedding for extracting features from social relational content to infer users' age, gender and personality traits of social media users. Forth, we empirically evaluate UDMF for the task of user profiling in social media and compare its performance with that of state-of-the-art methods on a sample dataset with 5K users from Facebook.

WSDM'18, , February 5–9, 2018, Marina Del Rey, CA, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5581-0/18/02…$15.00
https://doi.org/10.1145/3159652.3159691

2 RELATED WORK

Recently, modeling heterogeneous data sources has gathered significant interest. Integrating two or more sources of data to form a unified picture or make a better decision are the main goals of data integration frameworks. Modeling of different data sources and modalities provides a benefit for various multimedia tasks in sensor networks, robotics, and video and image processing. In addition, integrating multiple media such as textual data with audio and video content has been successful in various applications such as emotion detection [21], detecting events from sports videos [33], and wearable robotics by sensor data [16]. In this paper, we design a hybrid model for user profiling which incorporates both feature level and decision level of combining users' data in social media.

Many techniques have been proposed for modeling multimodal data sources. For an overview, we refer to [2]. In this section, we first discuss related work in user profiling in social media and then discuss existing work in modeling multiple data sources using deep neural networks.

User Profiling: There is a substantial body of existing work on automatically inferring a user's characteristics from the user's digital footprint in social media platforms. Existing single-source models usually leverage either only text, images, or relations.

Machine learning models have been trained to infer the age, gender, and personality traits of users based on the *textual context* they produce, including blog posts and status updates [6, 22, 24]. Author profiling has gained a lot of attention in the past few years. Workshops and competitions such as PAN[1] that focus on various features and techniques to predict age and gender of authors, or shared tasks such as WCPR[2] for personality prediction are a few examples of recent efforts.

Independently of this, recently important progress has been made on age and gender identification from *visual content* using deep neural networks. Rothe et al. for instance successfully used a Convolutional Neural Network (CNN) framework to detect the age and gender of users from their face [23]. There are competitions concentrating on this task as well, such as the LAP Challenge 2016 on predicting apparent age estimation and gender classification of images[3]. Less work has been done on inferring personality traits from visual content. In [3], Biel and Gatica-Perez predict the personality of Vloggers (YouTube bloggers) based on their visual and audio content. Identifying personality traits from a static image such as a profile picture is mostly uncharted territory. Recently, in [12], facial features (i.e., Face++ features) are extracted from Twitter profile pictures to predict users' personality. In this paper we use a similar approach, based on the Oxford project features (see Section 4.1) which is consumed by our heterogeneous user profiling model as one of the users' data sources.

Existing work on inferring user characteristics from *relational content* focuses typically either on using homophily or heterophily relations among friends [5, 14], or indirect relations among users such as shared Facebook page likes [11]. In this paper, we use a novel embedding called Node2Vec, in which we extract relational features from the graph using a deep neural network architecture.

Although much progress has been made in the area of user profiling, leveraging multimodal information to this end is a largely unexplored area of research. Most of the related work integrates data sources at the feature level [34]. The closest work to ours is the research by Wei et al. [29], who used a framework to integrate textual data, avatars (i.e., visual data) and responsive patterns of social media users using an ensemble method. They leveraged neural networks in their work by extracting features using a CNN architecture. The work presented in this paper is different in the following ways: (1) we propose a hybrid framework which outperforms an ensemble method such as the one proposed in [29]; and (2) we infer age and gender of users in addition to predicting personality traits of all users. Wei et al. only predict personality traits of the extreme users and remove more than 70% of the users with neutral personality from the inference.

Multimodal Modeling in Deep Neural Networks: There is a large body of work on integrating multimodal data sources in deep neural networks. Our framework utilizes stacking and power-set combination for a hybrid integration of user data in social media (see Section 3). In contrast, most of the related works either combine the data sources at the feature level or at the decision level. For instance in [15], the authors propose a deep autoencoder network to learn a multimodal feature representation for the task of audio-visual speech recognition. They pre-trained their deep autoencoder network using sparse Restricted Boltzmann Machines (RBM). Hybrid integration of data sources has been studied as well, however without exploring every combination of the data sources. For instance in [31], the authors propose a hybrid deep learning framework for video classification, in which they combine CNN features with a Long Short Term Memory (LSTM) network. The features extracted from CNNs are combined using a regularized feature level integration network, and LSTM is used on the temporal modality. A hybrid architecture automatically combines the LSTM and CNN features.

The idea of stacking in deep neural networks has been used in various ways. For instance, in [30], the authors proposed to stack so-called bottleneck features (i.e., vectors consisting of the activations at a bottleneck layer with a small number of hidden units compared to the other hidden layers in the network) for the task of speech recognition. In [13], the authors use stacking to stack two deep neural networks. The first network is an unsupervised network to extract features and the second network is a supervised network for the prediction task. In this paper, we use stacking for multi-label prediction, in which we stack the output of the networks which are trained per each target label (see Section 3). In addition, we also stack two deep neural networks (i.e., supervised and unsupervised) similar to [13] to extract features from a relational graph with the novel deep neural network architecture "Node2Vec" for the task of user profiling (see Section 4). A recent related work in combining features at the feature level is [32] in which the authors propose attentional factorization machines to learn the weight of feature interactions. The authors consider a pair-wise interaction layer to learn the interaction between features after the embedding layer, while in this paper we propose a power-set combination of the data sources at the embedding layer which considers all the possible combinations of features.

3 UDMF: USER PROFILING THROUGH DEEP MULTIMODAL FUSION

The main goal of modeling multiple data sources is to integrate two or more sources of data/knowledge and create a single representation that provides a more accurate description of the data sources than any of the individual ones. To design such a framework, one of the main considerations is the level where the integration of data sources happens. There are two widely known strategies, namely the *early approach* and the *late approach*. Figure 1 sketches the early and late approaches in a deep neural network architecture.

[1] http://pan.webis.de/

[2] https://sites.google.com/site/wcprst/home/wcpr14

[3] http://gesture.chalearn.org/

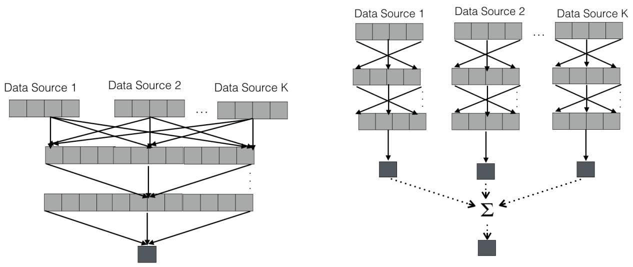

Figure 1: Early and late approaches of integrating multiple data sources in a deep neural network architecture.

The early approach is the strategy of integrating the data sources at the feature level. One of the main advantages of using the early approach is that correlation among different data sources and modalities are taken into account. The correlation among different modalities represents how different sources of knowledge co-vary with the other one. These types of correlation between data sources can provide additional cues in the integration process. But, different data sources and modalities do not necessarily correlate with each other. Therefore, in addition to using the dependency among data sources, it is often useful to fuse independent modalities to obtain a better decision. Let us consider the case of user profiling in social media. In this case, multiple modalities such as users' post and pictures can be used as a means of interaction with the platform. It is sometimes very hard to fuse these modalities at the feature level due to a lack of direct correspondence between their features.

The other popular approach of modeling multiple data sources is the late approach, where integration happens at the decision level. For instance, a linear weighted combination is the simplest decision (late) integration technique used. The widely used majority voting ensemble approach is a special case of this.

By incorporating both levels of modeling multiple data sources in our user profiling framework, we are able to take advantage of both approaches. Our hybrid model has two main properties. First, it leverages all sources of users' data and incorporates the correlation between modalities by mapping all combinations of data sources into shared representations. Second, in our model, integration of data sources also happens at the decision level when we combine the decision of all combinations of data sources. Moreover, to integrate the correlation among the target variables in a multi-task learning setup, we iteratively utilize the decision of our data integration framework for dependent tasks in the learning process.

We choose deep neural networks to implement our user profiling model for several reasons. First, it is easy to combine various data sources by using a shared representation between modalities. Second, we are able to combine data sources with non-linear functions which has been proved to enhance the learning process. Third, we are able to use neural networks on raw data sources and extract features using unsupervised approaches, i.e., the Node2Vec embedding that we use in Section 4. In the rest of this section, we introduce the structure for connecting data sources to the neurons in our proposed user profiling model, UDMF. To integrate data sources in UDMF, we design two mechanisms: *stacking* and *power-set combination*. To present them, we start from a general setting for a multilayer feedforward network described by an acyclic graph. We begin with a single data source D as input. The degree of activation of unit i on layer h is computed as:

$$U_i^h(D) = f(\sum_j w_{ij}^{hl} \cdot U_j^l(D)) \tag{1}$$

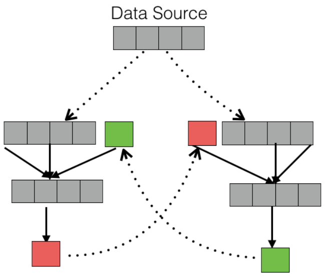

Figure 2: Stacking of 2 target variables given one data source.

where l is the layer that proceeds layer h, and j ranges over all the neurons on the layer l connected to layer h. w_{ij}^{hl} is the weight of the connection between neuron j on layer l and neuron i on layer h. f is an activation function which can be a non-linear function such as sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ for the output layer or ReLU ($ReLU(x) = \max(x, 0)$) for the hidden layers.

The degree of activation of unit i on layer 0, i.e., the layer connected to the input data source, is defined as follows (where $h = 0$):

$$U_i^0(D) = f(\sum_j w_{ij} \cdot D_j) \tag{2}$$

with w_{ij} representing the weight on the edge from input neuron j to neuron i on layer 0, and D_j denoting the input neuron j of the data source D, where j ranges over all values from 0 to the size of the input data source D, $|D|$.

Stacking. The stacking mechanism that we introduce in this paper makes the UDMF framework suitable for multi-task learning where target variables are correlated with each other. In a user profiling set up – which is the main focus of this paper – user attributes are correlated with each other: for instance inferring users' age becomes easier if we know their gender, and similarly, as users' demographics and personality traits are correlated, predicting one helps in predicting the other one. Farnadi et al. in [6] discuss the advantage of using multi-task learning in predicting users' personality traits using three social media data sets.

Figure 2 demonstrates stacking of two target variables given one data source in which two similar networks are trained per each target variable, but the input of each network consists of the input data source and the predicted output of the other target variable.

Assuming multiple epochs in learning the neural networks, Equation 2 is replaced with Equation 3, in which z ranges over the target variables. The degree of activation of unit i on layer 0 at epoch q is of the form:

$$U_i^{0q}(D) = f(\sum_j w_{ij} \cdot D_j + \sum_z w_{iz} \cdot \alpha_z \cdot t_z^{q-1}) \tag{3}$$

where α_z is a gating 0-1 variable. If z is equal to the target variable of the network, the value of $\alpha_z = 0$, otherwise $\alpha_z = 1$. In this way, the network for a specific target variable takes the predicted values t_z^{q-1} of the target variables of the other networks made during epoch $q - 1$ as input. We initialize the value of the target neurons where $q = 0$ with zero, hence:

$$U_i^{00}(D) = f(\sum_j w_{ij} \cdot D_j) \tag{4}$$

At each epoch the predicted values of the target variables are updated based on the previously predicted values of the other target variables. As shown in Figure 2, for a sample configuration with two target variables, for each target variable, we create a network with a similar architecture. At each epoch, the predicted value of each target variable is stacked as an input for the other network. This configuration can easily be extended to more than two networks, i.e., more than two target variables. We can also update the predicted values of the target variables at every ten epochs instead of every epoch. In the experiments presented in section 4, we update the predicted values every 10 epochs and iterate 10 times to get 100 epochs.

Power-set Combination. Let $DS = \{D_1, D_2, \ldots, D_k\}$ be the finite set of k data sources that we want to integrate. Note that k is a small number as k represents the number of data sources that we have per each user. In social media k is typically between two to at most five sources of user data (i.e., textual, visual, relational, temporal, geo location). In our proposed power-set combination approach, we incorporate correlations among features and data sources by an early integration approach of all subsets of DS. We then combine their predicted outcome as a late integration approach with an ensemble method. Therefore, the UDMF model is a hybrid data integration model. The input layer of the UDMF framework consists of inputs originating from 1 to k data sources. Each neuron in the first hidden layer connected to the input layer can potentially be connected to the input neurons of any subset of these data sources. Precisely, given the set DS of the k data sources, we calculate the power-set of DS, i.e. the set of all subsets of DS. We exclude the empty set and therefore the number of subsets of DS under consideration is $2^k - 1$. Per each non-empty subset $\mathcal{D}$ in the power-set of DS we build a mini-DNN. The activation level of neuron i on layer 0 of each mini-DNN which combines $|\mathcal{D}|$ data sources at epoch q is computed as:

$$U_i^{0q}(\mathcal{D}) = f(\sum_{D \in \mathcal{D}} \sum_{j} w_{ij} \cdot D_j + \sum_{z} w_{iz} \cdot \alpha_z \cdot t_z^{q-1}) \qquad (5)$$

where $\mathcal{D} \in \mathcal{P}(DS)$ is the subset of the input data sources leveraged in the mini-DNN. Equation (5) is the counterpart of Equation (2) in the UDMF framework.

As an example of power-set combination of the data sources in UDMF, we assume two available data sources A and B, $DS = \{A, B\}$, so the power-set $\mathcal{P}(DS)$ is $\{\{A\}, \{B\}, \{A, B\}, \{\}\}$. Therefore, we can make three mini-DNNs where A, B and combined A and B make up the respective input layers. We present the UDMF network for these two data sources and two target variables in Figure 3. As shown in the figure, we train three mini-DNNs for each target variable, therefore in total we train six mini-DNNs. The output of each mini-DNN is stacked as input to the sister mini-DNNs at the end of each epoch for the training at the next epoch.

If we have three data sources. i.e., textual, visual and relational, and p target variables (e.g., $p = 7$, see Section 4), we need to train $(2^3 - 1) \cdot 7 = 49$ mini-DNNs. As we stacked the target variables, we have an inter-connected network of seven mini-DNNs per each data source combination from the power-set of the data sources $\mathcal{P}(DS)$. Hence, we have seven DNN models that combine the data sources in various ways and each one of them includes seven mini-DNNs which are inter-connected to each other. Since modeling multiple data sources at the decision level of these seven DNN models happens in UDMF by majority voting as a late integration step, each multi-target network can be trained separately from each other. Training each power-set combination multi-target network in parallel reduces the time to train the UDMF to a great extent.

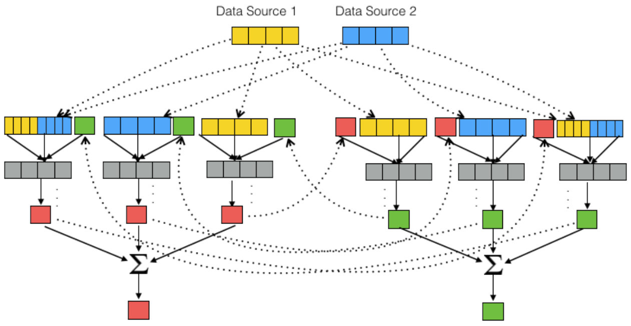

Figure 3: The architecture of UDMF with stacking of 2 target variables and power-set combination of two data sources.

4 EVALUATION

We train and test the UDMF framework with a subset of the MyPersonality project dataset[4]. MyPersonality was a popular Facebook application introduced in 2007 in which users took a standard Big Five Factor Model psychometric questionnaire [7] and gave consent to record their responses and Facebook profile. The dataset contains information about each user's demographics, friendship links, Facebook activities (e.g., number of group affiliations, page likes, education and work history), status updates, profile picture and Big Five Personality scores. However, not all of this information is available for all users. We selected users who mention English as their language, and who provide age, gender, personality, status updates, page likes and a profile picture. To increase the chance that the image depicts the profile owner, we first selected profile pictures with only one face using the Project Oxford Face detector API[5]. By removing the Facebook pages with less than 3 likes by users in this dataset, our final dataset includes 49,372 pages, and 724,948 page like relations for 5,670 users.

Personality traits are commonly described using five dimensions (known as the Big Five), i.e., Extraversion (Ext), Agreeableness (Agr), Conscientiousness (Con), Neuroticism (Neu), and Openness (Opn). The range of the personality scores in this dataset is between $[1, 5]$. We use the median value to create binary classes for each characteristic, where the median value for age = 23, Opn = 4, Con = 3.5, Ext = 3.5, Agr = 3.65, and Neu = 2.75. We evaluate the proposed user profiling model for the tasks of predicting age, gender and personality traits of Facebook users using their textual (status updates), visual (profile picture) and relational data (page likes).

We systematically perform 10-fold cross-validation. Since all the characteristics that we aim to predict are binary (i.e., positive class vs. negative class), to evaluate the results, we use AUC scores. AUC is the area under the ROC curve (i.e., receiver operating characteristic), which is created by plotting the true positive rate (i.e., portion of positives that are correctly predicted as such) against the false positive rate (i.e., portion of negatives that are wrongly predicted as positive). The natural language processing, machine learning and deep learning techniques in the following subsections are implemented using the scikit-learn[6] and keras[7] libraries in Python.

[4] http://mypersonality.org/

[5] https://www.microsoft.com/cognitive-services/en-us/face-api

[6] http://scikit-learn.org/

[7] https://keras.io/

To be able to correctly measure the effect of UDMF in integrating various data sources, as the basic building blocks of our configurations, we design simple DNNs consisting of three layers, where the first layer is the input layer, the second layer is a hidden layer which has 100 neurons per each data source as input and the last layer is a sigmoid layer that represents the result. For all the DNNs, the hidden layer of all the networks that we compare leverage ReLU as the activation function to model a non-linear combination of the inputs. We used Adam as the optimization algorithm, and we train all DNN models in this paper for 100 epochs, with a batch size set to 128. The other parameters of the DNN models are set by default values. We compare the performance of the models with the simple majority baseline algorithm that assigns the majority class from the training instances to the test instances. In addition, we compare with baseline methods that best learns combinations of modality features (early fusion)/decisions (late fusion) with the training data to show the power of the UDMF approach.

4.1 Data Source Embeddings

In practice, data always contain noise and we cannot expect to arrive at a good data representation without data cleaning and preprocessing. Which data processing to use under which constraints depends very much on the type of application. Before showing how UDMF fuses users' data, in this section we discuss how we represent each data source for the task of user profiling in social media. We define three data source embeddings: a data source embedding from the textual content, a data source embedding from the visual content, and a data source embedding from the relational content. We obtain the data source embeddings using the dataset described above.

Textual data source embedding: To build the textual data source embedding, we combine the status updates of each user in the dataset into one document per user. We represent each user with 88 Linguistic Inquiry and Word Count (LIWC) [19] features extracted from her/his status updates, consisting of features related to (a) standard counts (e.g., word count), (b) psychological processes (e.g., the number of anger words such as *hate, annoyed, . . .* in the text), (c) relativity (e.g., the number of verbs in the future tense), (d) personal concerns (e.g., the number of words that refer to occupation such as *job, majors, . . .*), and (e) linguistic dimensions (e.g., the number of swear words). For a complete overview, we refer to [26]. We compared the performance of using various feature sets, namely, LIWC, n-grams (n=1, 2, 3), a 300-dimensional pre-trained GloVe [20] vector based on Twitter data, and a 300-dimensional pre-trained fastText [10] vector based on English Wikipedia data with default parameters. The DNN models with LIWC features as the input layer significantly outperform similar DNN models with other feature sets and embeddings as the input layer, therefore in the rest of the paper we use LIWC features as our textual data source embedding. Results of the LIWC-based DNN models are presented in Table 1 as "Text". Due to the space restriction, we omit the results of DNN models based on other representations of the text from the paper.

Visual data source embedding: For each user we use his/her profile picture and extract 64 facial features using the Oxford Face API [4]. The extracted features are face rectangle features to capture the location of the face in the image, face landmark features which include 27-point face landmarks pointing to the important positions of face components, face characteristics including age, gender, facial hair, smile, head position and glasses type. We compared the performance of the Oxford features as the input layer, with a 128-dimensional activation vector extracted from the last layer (before the softmax) of the pre-trained VGG-16 and VGG-19 models [25] on ImageNet. The DNNs with the Oxford features as the input layer significantly outperform the VGG-based models specifically for the task of age and gender prediction. Results of the Oxford-based DNN models are presented in Table 1 as "Image". Due to the space restriction, we omit the results of DNN models based on other embeddings of the profile pictures from the paper.

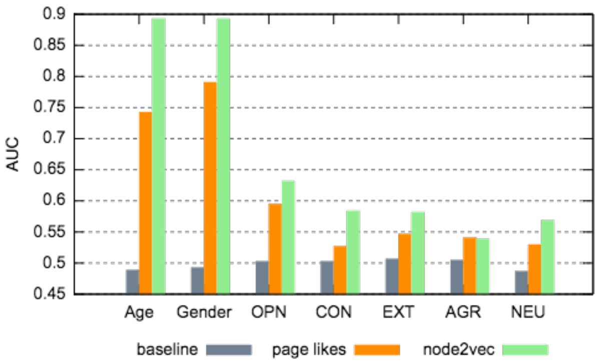

Figure 4: Node2Vec features extracted from users' page likes outperform using only pages that users like as features for the tasks of inferring gender, age, and Big Five personality traits.

Relational data source embedding: To represent users with pages that they like, we train an unsupervised deep neural network approach called Node2Vec on our relational graph [1]. Node2Vec is extending the Skip-gram architecture [8] to networks. Let $G = (V, E)$ be our relational graph, and f be the mapping function to represent nodes with features (i.e., $f : V \to \mathbb{R}^d$). The Node2Vec model optimizes the following objective function that maximizes the log-probability of observing a network neighborhood $NS(u)$ for a node u conditioned on its feature representation, given by f:

$$\max_{f} \sum_{u \in V} \log Pr(NS(u)|f) \tag{6}$$

We learn a mapping of users to a low-dimensional space of features that maximizes the likelihood of preserving network neighborhoods of users and pages. To this end, we train a Node2Vec model using our page like relations. Using features extracted from the Node2Vec model, we not only represent users with pages that they like (i.e., their neighbors), but also we find similar users by a flexible biased random walk procedure (Node2Vec walks) to produce $NS(u)$ that can explore neighborhoods in both a Breadth-First Sampling (BFS) as well as a Depth-First Sampling (DFS) fashion. BFS samples nodes which are immediate neighbors of the source, while DFS samples nodes at increasing distances from the source node. We iteratively perform the random Node2Vec walk on the graph to sample nearest neighbors for each node and then train a Skip-gram architecture to find embeddings for each node. We set the number of dimensions d to 127. From the output embedding, we select embeddings of the nodes which represent users in our domain and ignore the representation of the pages. The models using the Node2Vec features have outperformed state-of-the-art techniques on multi-label classification and link prediction in several real-world networks [1]. To the best of our knowledge, we are the first to use the Node2Vec embedding for user profiling using social relational content in social media.

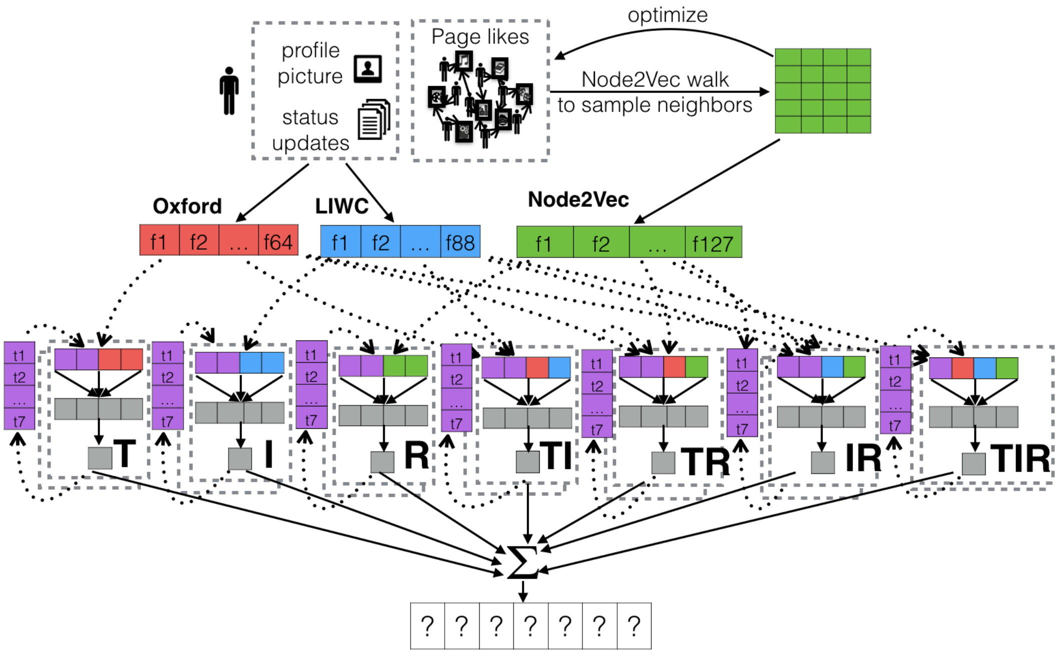

Figure 5: The architecture of user profiling with UDMF

This new representation of users and pages in a new space, allows us to gain additional knowledge compared to an adjacency matrix where users (i.e., rows) are represented by pages that they like (i.e., columns). To investigate the performance of a model using the Node2Vec embeddings for the user profiling task, we compared the performance of using the Node2Vec feature representation with the model proposed in [11]. We call the latter model the *page likes model* in the rest of this paper. In the page likes model, each row represents a user in the dataset and columns represent pages. The value of each matrix entry is one if the user likes that page in the dataset, otherwise it is zero. In [11], Lasso is used to predict the Big Five personality traits, however since our labels are binary, we use ridge regression which is a linear least squares classifier with $l2$ regularization. We set all parameters to their default values from the scikit-learn library.

The results are presented in Figure 4. We compared the Node2Vec model (shown with green bars as Node2Vec) with the majority baseline (shown with grey bars as baseline) and the page likes model (shown with orange bars as page likes) where both the Node2Vec and page likes models outperform the majority baseline and the Node2Vec models outperform the page like models in predicting all labels. Specifically, the Node2Vec models for age and gender predictions yield an AUC score which is close to 0.9. Note that we use very simple DNN models with only three layers to collect these results, however using more layers, and/or using a regularizer may further increase the accuracy of the prediction and improve these results, but this is out of the scope of this paper.

4.2 Data Source Integration

To integrate data sources in UDMF, we design a mini-DNNs architecture, with one layer for the input neurons, one hidden layer which is fully connected to the input neurons and one output layer which has the output of the learning task using the sigmoid activation function. We make similar mini-DNNs per each target variable that are integrated in the stacking process. In this way, for each subset of data sources, we have p mini-DNNs of this type which are getting updated at each epoch from the output of $p - 1$ other mini-DNNs. To fuse the final result, we apply majority voting to determine the label.

In a user profiling setting, each unlabeled user is assigned labels from a finite set of labels, e.g., in our case $l = \{$ *female*, *young*, *Opn*, *Con*, *Ext*, *Agr*, *Neu*$\}$. The architecture of our user profiling framework with UDMF is shown in Figure 5. We stack $p = 7$ similar mini-DNNs where each network is trained for one target variable from a set of labels l. The parameters of each network can be tuned for the task and can possibly be different from the other networks. In this paper, however, for the sake of simplicity, we choose a common neural network architecture representation with the same choices across all the target variables to fairly compare the results with each other.

The different sources of information (profile picture, status updates, page likes) can all contribute to the construction of an accurate user profile. With regard to text, we observe that older users post more greetings in their status updates (e.g., "Happy birthday, happy christmas"), Neurotic users use more swear words, and Conscientious users utilize more terms related to time. The presence or absence of facial hair, such as a beard or mustache, is the best indicator for users' gender, and head position is a good indicator for Extraversion. Pages that users like provide useful clues as well, e.g. "I Love Being A Mum!" is a good indicator of a user's age and gender[8].

5 EXPERIMENTAL RESULTS

In this section, we evaluate the UDMF for the task of inferring users' age, gender and personality traits in Facebook. UDMF involves

[8] Note that in this paper, we only use the user-page like relations, and not the actual titles of the pages that users like on Facebook.

Table 1: Mean and standard deviation of area under the curve (AUC) scores in inferring age, gender and personality traits with one, two and three data source embeddings using mini-DNNs. For each category, results of using stacking (Equation 3) are shown with ✓, and results of not using stacking (Equation 2) are shown with ✗. All results are averaged over a 10-fold CV. In each column, the highest results are typeset in bold.

Model	Stack	Age	Gender	Opn	Con	Ext	Agr	Neu
Baseline		0.488	0.492	0.502	0.502	0.506	0.506	0.486
				One source				
Text	✗	0.741±0.022	0.668±0.020	0.550±0.016	0.575±0.017	0.536±0.016	0.547±0.016	0.523±0.016
	✓	0.748±0.022	0.668±0.020	0.553±0.017	0.574±0.017	0.545±0.016	0.550±0.016	0.524±0.016
Image	✗	0.552±0.016	0.915±0.027	0.502±0.015	0.500±0.015	0.504±0.015	0.512±0.015	0.520±0.016
	✓	0.550±0.016	0.897±0.027	0.516±0.015	0.511±0.015	0.518±0.015	0.519±0.015	0.541±0.016
Relation	✗	0.875±0.026	0.886±0.027	0.601±0.018	0.571±0.017	0.567±0.017	0.525±0.016	0.558±0.017
	✓	0.893±0.027	0.898±0.027	0.622±0.018	0.589±0.018	0.573±0.017	0.533±0.016	0.563±0.016
				Two sources				
Early approach	✗	0.734±0.022	0.873±0.026	0.569±0.017	0.588±0.018	0.536±0.016	0.545±0.016	0.547±0.016
TI	✓	0.746±0.022	0.864±0.026	0.546±0.016	0.568±0.017	0.542±0.016	0.546±0.016	0.536±0.016
Early approach	✗	0.878±0.026	0.896±0.027	0.610±0.018	0.586±0.018	0.567±0.017	0.535±0.016	0.554±0.017
TR	✓	0.891±0.027	0.899±0.027	0.627±0.019	0.601±0.019	0.572±0.017	0.551±0.016	**0.574±0.017**
Early approach	✗	0.878±0.026	0.951±0.028	0.606±0.018	0.574±0.017	0.569±0.017	0.524±0.016	0.562±0.017
IR	✓	0.895±0.027	0.951±0.028	0.633±0.019	0.592±0.018	0.577±0.017	0.537±0.016	0.564±0.017
				Three sources				
Ensemble	✗	0.876±0.026	**0.952±0.028**	0.603±0.018	0.587±0.018	0.569±0.017	0.537±0.016	0.562±0.017
(Late approach)	✓	0.893±0.027	0.949±0.028	0.626±0.019	0.606±0.018	**0.582±0.017**	0.549±0.016	0.570±0.017
Early approach	✗	0.887±0.027	0.947±0.028	0.617±0.018	0.577±0.017	0.567±0.017	0.541±0.016	0.566±0.017
TIR	✓	**0.899±0.027**	0.934±0.028	**0.635±0.019**	**0.607±0.018**	0.560±0.018	**0.551±0.016**	0.572±0.017

two main strategies: stacking and power-set combination. We first examine the output of the UDMF framework using only stacking with a single source, a combination of two sources, and finally all three sources (Table 1). Then, we examine the capabilities of our hybrid UDMF framework with both stacking and power-set combination of two and three data sources in modeling multiple data sources (Table 2). Except for the parameter being tested, all other parameters assume default values for all the models. The learning curve of training the mini-DNNs with three data sources (i.e., shown as TIR in Figure 5) for the case of age prediction has been shown in Figure 6. As shown in the figure, all networks converge after 100 epochs. We omit the learning curve of other mini-DNNs networks, and other traits due to their similar behavior.

Unimodal baselines: To evaluate UDMF, we first get results using only one of the data sources with the basic DNN structure that we presented above (using Equation 2). These results are presented in the first rows for Text, Image, and Relation in Table 1. As expected, using Oxford features extracted from the profile picture of users to build the image model outperforms the majority baseline in predicting age and gender, however for personality traits these features perform poorly. The text model using the LIWC features on the other hand outperforms the majority baseline for predicting all the labels, and the model using the Node2Vec features performs reasonably well in inferring all the traits. The best performing model which uses only a single source of data is the relation model using data source embeddings with Node2Vec features.

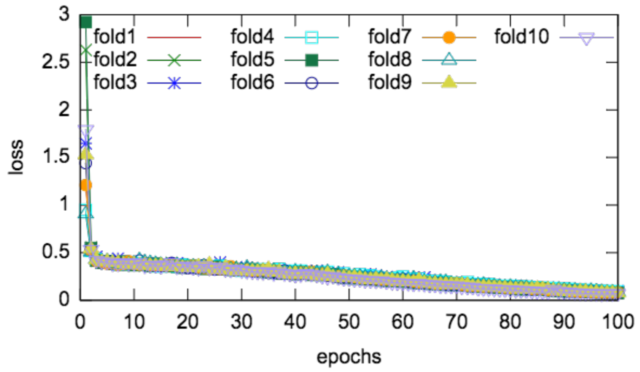

Figure 6: The training learning curve of infering age in all folds in mini-DNNs called TIR, where all three data sources are combined

Multimodal baselines: We then combine the data sources both at feature level and decision level. We fuse all combinations of two sources and three sources of data. The results of combining two sources at the feature level are presented in the odd rows of the second section (i.e., Two sources) in Table 1, where *TI* is the combination of the textual and visual data sources, *TR* indicates the combination of the textual and relational data sources and *IR* refers to the combination of the visual and relational data sources. As expected, integrating data sources at the feature level (i.e., in an early integration manner) tends to performs better than the single source models in the first section of the table. The results of the early and late integration approach in combining three sources of data are close to each other but the early fusion approach works slightly better than the late integration approach (i.e., majority voting). These results are presented in the odd rows of the last section (i.e., Three sources) of Table 1.

To examine the stacking property, we train all networks with one, two and three data sources with stacking enabled (using Equation 3). The results are presented in the even rows of the Table 1 below the results of not using stacking per each category, where the "Stack" option is marked as enabled with ✓. It is interesting that training a network with stacking outperforms a similar network without stacking for most of the DNN networks that we examined either with single source DNNs and multi-source DNNs. Similarly, the networks trained with stacking outperform the networks without stacking both at the feature level and at the decision level. This indeed indicates the correlation among the target variables (i.e., age, gender and personality traits).

We then evaluate the UDMF architecture with both power-set combination and stacking while integrating two and three data sources (using Equation 5). As shown in Table 2 the AUC scores of

Table 2: Mean and standard deviation of AUC scores in inferring age, gender and personality traits by fusing two and three data sources in UDMF (Equation 5). All results are averaged over a 10-fold CV. In each column, the highest results are typeset in bold.

Model	Age	Gender	Opn	Con	Ext	Agr	Neu
One/Two sources							
Page likes	0.743±0.020	0.699±0.022	0.605±0.017	0.516±0.016	0.555±0.016	0.540±0.0161	0.527±0.016
LR (T)	0.711±0.021	0.654±0.020	0.564±0.017	0.568±0.017	0.551±0.016	0.548±0.016	0.530±0.016
LR (I)	0.584±0.017	0.858 ±0.026	0.514±0.015	0.520±0.015	0.528±0.016	0.528±0.016	0.525±0.016
LR(T,I)	0.711±0.017	0.852 ±0.025	0.555±0.017	0.564±0.017	0.551±0.016	0.550±0.016	0.542±0.016
UDMF(T,I)	0.756±0.023	0.886±0.027	0.569±0.017	0.575±0.017	0.552±0.017	0.552±0.016	0.539±0.016
UDMF(T,R)	0.879±0.026	0.943±0.028	0.628±0.019	0.607±0.018	0.580±0.017	**0.564±0.017**	0.575±0.017
UDMF(I,R)	0.892±0.027	0.955±0.029	0.630±0.019	0.607±0.018	0.587±0.018	0.551±0.016	0.571±0.017
Three sources							
Weighted Soft Voting	0.656±0.019	0.861±0.026	0.523±0.016	0.0523±0.016	0.508±0.015	0.507±0.015	0.518±0.015
Random Forest (100)(T,I,R)	0.786 ±0.023	0.900±0.027	0.588±0.018	0.564 ±0.017	0.544±0.016	0.549±0.016	0.538±0.016
LR(T,I,R)	0.808 ±0.024	0.888±0.027	0.603±0.018	0.585 ±0.018	0.550±0.017	0.550±0.016	0.572±0.017
UDMF(T,I,R)	**0.903±0.027**	**0.956±0.029**	**0.647±0.019**	**0.615±0.018**	**0.592±0.018**	0.556±0.017	**0.580±0.017**

the UDMF models are better than those obtained with early integration and late integration of the data sources as presented in Table 1. The UDMF framework in which we combined all three sources of data for user profiling outperforms all the competing networks presented in this paper for all tasks except for the Agreeableness class. In the case of Agreeableness, the UDMF model which integrates the textual and relational content outperforms the UDMF of all three sources of data. This result may be due to the poor performance of visual content in inferring the Agreeableness trait, producing noise instead of providing any additional evidence to fuse the information. We get highly accurate results in predicting age and gender of users with an AUC score of more than 0.90 in the case of age prediction and more than 0.95 for the task of gender prediction.

The UDMF framework outperforms state-of-the-art techniques for predicting age, gender and personality traits, as shown in the first section of Table 2. The first row corresponds to the *the page likes model* of Kosinski et al. [11]. The second row, containing results of logistic regression (LR) applied on LIWC features, is inspired by Farnadi et al. [6] who used decision trees and support vector machines. In our experiments, we found that LR outperformed these approaches, therefore we only include the results of using LIWC features with LR as a learner (shown as LR(T) in Table 2). Similarly we add the results of using Oxford features with LR (shown as LR(I)) and an early integration approach inspired by related works in combining features at the feature level to combine the textual and visual features (shown as LR(T,I)). In addition, we include an ensemble method and early integration approach using LR (shown as LR(T,I, R)), random forest with 100 estimators and a weighted soft voting approach of the results of the decision tree, support vector machine and LR. UDMF outperforms all these approaches.

We tested all the networks with various dropout parameters 0.1, 0.2, 0.3, 0.4, and 0.5. We find that the models are not very sensitive to variations of the dropout rate. The difference between using dropout and not using this procedure differs in a small amount, thus to keep the simplicity of the networks, we omit the results from this paper. Note that we chose a simple architecture for designing the mini-DNN networks, and we used a similar DNN architecture throughout this paper for fair comparison, however one could design a more sophisticated DNN for the task of user profiling to enhance the learning power of UDMF.

6 CONCLUSION AND FUTURE DIRECTIONS

In this paper, we have introduced a hybrid user profiling architecture in deep neural networks which we call UDMF. UDMF has two simple and yet effective properties, namely stacking and power-set combination strategies, to better integrate users' data in social media for multi-target learning task. UDMF combines different modalities both at the feature level and decision level to predict accurate multiple attributes of social media users given their user generated content and social relational content. We evaluated the UDMF architecture for the user modeling task on more than 5K users from Facebook. We built three data sources from users' textual, visual and relational content given their status updates, profile picture and pages that they like on Facebook. The results showed how stacking and power-set combination in UDMF enhance the learning power in combining the data sources. To make a user profile, we predicted age, gender and personality traits of users with UDMF. We obtained highly accurate results, including an AUC score of more than 0.9 for the task of age prediction and 0.95 for the task of gender prediction.

In this paper, we trained our user profiling UDMF framework on a sample of data from Facebook. Due to the small size of the textual and visual data, we could not use a neural network architecture to learn features. Another interesting direction for future work is to use larger datasets to learn features while integrating them with UDMF. Moreover, integrating other sources of users' data such as temporal data and geo-location data such as users' check-in information, remains as a future direction of this work.

With regard to the data integration part of UDMF, there are two main directions for future work. First, data integration frameworks are either exploiting overlap between modalities to "reinforce the signal", or gathering different pieces of information from different sources to obtain a full picture. An ideal user profiling framework utilizes each source of users' data/knowledge to discover part of the information, and therefore by integrating multiple sources of data, a data integration framework should put all pieces of data together and provide a descriptive and comprehensive representation of users in social media. One could extend the UDMF framework with a new objective function that learns the degree to which integrating modalities share evidences and the degree to which data sources provide complementarity of evidence that is captured in the learned representations.

Second, another important factor in modeling mutiple data sources is that different modalities may have varying capabilities of accomplishing a specific task. For example, we observed that a profile picture is not a suitable source of data to infer a user's personality traits by itself, however, combining this source with other sources of data such as textual content and relational content boost the performance of the learner. Although UDMF performs well in combining strong and poor sources of data, it is an interesting direction for future work to incorporate the confidence level of each source in the integration process. Learning a confidence level is a complex task that may change by considering various factors such as noise in the data. Designing an approach for learning this with a deep neural network is an open path to explore.

REFERENCES

[1] node2vec: Scalable feature learning for networks.

[2] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[3] J.-I. Biel and D. Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Proc. of IEEE Transactions on Multimedia*, 15(1):41–55, 2013.

[4] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2707–2714. IEEE, 2010.

[5] G. Farnadi, Z. Mahdavifar, I. Keller, J. Nelson, A. Teredesai, M.-F. Moens, and M. De Cock. Scalable adaptive label propagation in Grappa. In *Proc. of IEEE International Conference on Big Data*, pages 1485–1491, 2015.

[6] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, and M. De Cock. Computational personality recognition in social media. *User Modeling and User Adapted Interaction*, pages 1–34, 2016.

[7] L.-R. Goldberg, J.-A. Johnson, H.-W. Eber, R. Hogan, M.-C. Ashton, C.-R. Cloninger, and H.-G. Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96, 2006.

[8] Y. Goldberg and O. Levy. Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[9] G. I., B. Y., and C. A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[10] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[11] M. Kosinski, D. J. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proc. of the National Academy Of Sciences (PNAS)*, 110:5802–5805, 2013.

[12] L. Liu, D. Preotiuc-Pietro, Z. Riahi Samani, M. E. Moghaddam, and L. Ungar. Analyzing personality through social media profile picture choice. In *Proc. of the International AAAI Conference on Web and Social Media*, 2016.

[13] A. B.-J. Low, C-Y.and Teoh. Stacking-based deep neural network: Deep analytic network on convolutional spectral histogram features. *arXiv preprint arXiv:1703.01396*, 2017.

[14] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444, 2001.

[15] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proc. of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011.

[16] D. Novak and R. Riener. A survey of sensor fusion methods in wearable robotics. *Robotics and Autonomous Systems*, 73:155–170, 2015.

[17] S. Nowson and J. Oberlander. The identity of bloggers: Openness and gender in personal weblogs. In *Proc. of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 163–167, 2006.

[18] R. D. Oliveira, A. Karatzoglou, P. C. Cerezo, A. A. L. D. Vicuña, and N. Oliver. Towards a psychographic user model from mobile phone usage. In *Proc. of the International Conference on Human Factors in Computing Systems, CHI*, pages 2191–2196, 2011.

[19] J.-W. Pennebaker and L.-A. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77:1296–1312, 1999.

[20] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 14, pages 1532–1543, 2014.

[21] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.

[22] F. Rangel, P. Rosso, M. Potthast, B. Stein, and W. Daelemans. Overview of the 3rd Author Profiling Task at PAN 2015. In *Proc. of CLEF*, 2015.

[23] R. Rothe, R. Timofte, and L. Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proc. of ICCV, ChaLearn Looking at People workshop*, 2015.

[24] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] Y. R. Tausczik and J. W. Pennebaker. The Psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24–54, 2010.

[27] M. Tkalcic and L. Chen. Personality and recommender systems. In *Recommender Systems Handbook*, pages 715–739. Springer, 2015.

[28] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(3):328–339, 1989.

[29] H. Wei, F. Zhang, N. J. Yuan, C. Cao, H. Fu, X. Xie, Y. Rui, and W.-Y. Ma. Beyond the words: Predicting user personality from heterogeneous information. In *Proc. of the Tenth ACM International Conference on Web Search and Data Mining*, pages 305–314, 2017.

[30] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4460–4464, 2015.

[31] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proc. of the 23rd ACM international conference on Multimedia*, pages 461–470, 2015.

[32] J. Xiao, H. Ye, X. He, H. Zhang, F. Wu, and T.-S. Chua. Attentional factorization machines: Learning the weight of feature interactions via attention networks. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[33] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang. Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on Multimedia*, 10(7):1342–1355, 2008.

[34] Q. Zhu, M.-C. Yeh, and K.-T. Cheng. Multimodal fusion using learned text concepts for image categorization. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 211–220. ACM, 2006.